Proceedings of the 30th International Cosmic Ray Conference
Rogelio Caballero, Juan Carlos D'Olivo, Gustavo Medina-Tanco,
Lukas Nellen, Federico A. Sánchez, José F. Valdés-Galicia (eds.)
Universidad Nacional Autónoma de México,
Mexico City, Mexico, 2008

Vol. 4 (HE part 1), pages 809–812

30TH INTERNATIONAL COSMIC RAY CONFERENCE

The EM algorithm applied to the search of high energy neutrino sources

J. A. AGUILAR[1] AND J. J. HERNÁNDEZ-REY[1]
[1]*IFIC – Instituto de Física Corpuscular, CSIC – Universitat de València, apdo. 22085, E-46071 Valencia, Spain.*
aguilars@ific.uv.es

Abstract: The detection of astrophysical sources of high energy neutrinos is one of the most sought-after experimental goals in present astroparticle physics. The low number of signal events expected in high energy neutrino telescopes calls for powerful algorithms to disentangle clusters of small number of events from the background. In this contribution we explore the potentiality of the Expectation-Maximisation algorithm (EM) for the search of neutrino point-like sources in a generic kilometre-scale neutrino telescope located in the Mediterranean sea. The EM algorithm, widely used in clustering analysis, has been adapted in this work to the special case of low signal statistics in a relatively high background. We describe here how the problem has been tackled and compare our results to the well-known binning technique. The method can also be applied to other similar problems like for instance the search of nearby ultra-high energy cosmic-ray sources.

The Expectation-Maximisation algorithm

The Expectation-Maximisation algorithm [1] is a general approach to maximum likelihood estimation for finite mixture model problems. In these models different groups in the data are described by different density components, so that the total probability density function (pdf) can be expressed as $p(\mathbf{x}) = \sum_{j=1}^{g} \pi_j \, p(\mathbf{x};\boldsymbol{\theta}_j)$, where g is the number of mixture components, $\pi_j \geq 0$ are the mixing proportions that satisfy the unitary relation ($\sum_{j=1}^{g} \pi_j = 1$) and $p(\mathbf{x};\boldsymbol{\theta}_j)$, $j = 1, ..., g$, are the *component density functions* which depend on a parameter vector $\boldsymbol{\theta}_j$. The values of π_j and the components of $\boldsymbol{\theta}_j$ have to be found by maximising the likelihood. The number and form of the component density functions depend on the problem being tackled. Frequently, $p(\mathbf{x};\boldsymbol{\theta}_j)$ is taken to be the multivariate normal (Gaussian) density parametrised by a mean and a covariance matrix $\boldsymbol{\Sigma}_j$, ($\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$). In this particular case the equations given by the EM algorithm are well-known.

Having a data sample of n observations, the likelihood for a mixture model with g density components is given by $\mathcal{L}(\boldsymbol{\Psi}) = \prod_{i=1}^{n} \sum_{j=1}^{g} \pi_j p(\mathbf{x}_i;\boldsymbol{\theta}_j)$ where $\boldsymbol{\Psi}$ stands for the set of parameters $\{\pi_1, ..., \pi_g; \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_g\}$. In general it is not possible to solve explicitly $\partial\mathcal{L}/\partial\boldsymbol{\Psi} = 0$ and an iterative approach must be used. The EM algorithm is a general iterative procedure to maximise mixture model likelihoods. The idea is to assume that the set of observations given in the data make up a set of *incomplete* data vectors $\{\mathbf{x}\}$. The likelihood given by this *incomplete* data set can be expressed by $\mathcal{L}(\boldsymbol{\Psi}) = p(\{\mathbf{x}\}, \boldsymbol{\Psi})$. The unknown information missed in the data sample is as a matter of fact whether an observation belongs to a component or to another, in other words, we lack the information about the clustering structure of the population. Let $\{\mathbf{y}\}$ denote a *complete* data set, a version of $\{\mathbf{x}\}$ formed by a new vector $\mathbf{y}_i = (\mathbf{x}_i, \mathbf{z}_i)$, where $\mathbf{z}_i = (z_{i1}, ..., z_{ig})$ is simply a class indicator vector, that is, $z_{ik} = 1$, if $\mathbf{x}_i$ belongs to group k, and $z_{ik} = 0$, otherwise. With this *complete* data set, the total density function is:

$$g(\mathbf{y};\boldsymbol{\Psi}) = g(\mathbf{x},\mathbf{z};\boldsymbol{\Psi}) = p(\mathbf{x};\mathbf{z},\boldsymbol{\Psi}) f(\mathbf{z};\boldsymbol{\Psi}) \quad (1)$$

and the likelihood is $\mathcal{L}'(\boldsymbol{\Psi}) = g(\{\mathbf{y}\}, \boldsymbol{\Psi})$. We can obtain back the likelihood of the previous *incomplete* data set from $g(\{\mathbf{y}\}, \boldsymbol{\Psi})$ by integrating over

all possible $\{\mathbf{y}\}$ in which the set $\{\mathbf{x}\}$ is embedded, $\mathcal{L}(\mathbf{\Psi}) = p(\{\mathbf{x}\}, \mathbf{\Psi}) = \int \prod_{i=1}^{n} g(\mathbf{x}_i, \mathbf{z}; \mathbf{\Psi}) d\mathbf{z}$

The EM procedure consists in two main steps. In the Expectation step we evaluate the expectation of the complete data log-likelihood, $Q(\mathbf{\Psi}, \mathbf{\Psi}^{(m)})$, for the current value of the parameters estimate $\mathbf{\Psi}^{(m)}$. In the Maximisation-step, a new set of maximum estimate parameters $\mathbf{\Psi} = \{\mathbf{\Psi}^{(m+1)}\}$ are found by differentiating $Q(\mathbf{\Psi}, \mathbf{\Psi}^{(m)})$. Successive maximisations of the expected value $Q(\mathbf{\Psi}, \mathbf{\Psi}^{(m)})$ will lead to the maximisation of the desired *incomplete* likelihood. It can be shown that the EM algorithm has a general convergence property. For finite mixture models and assuming multivariate normal mixtures the two steps can be expressed in particularly simple forms suitable for code implementation.

Neutrino Telescope simulation

Results from the presently running experiments [2] and some theoretical predictions indicate that neutrino telescopes should have effective areas of about 1 km^2 or higher to make high energy neutrino astronomy. A new generation of kilometre-scale detectors are being designed or well under way [3]. Any gain in the power of the searching algorithms is of the utmost importance to claim the first hints of high energy neutrino source detection.

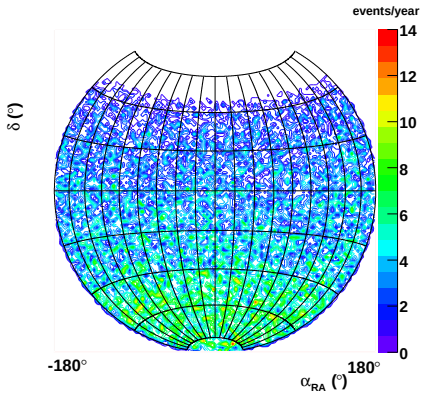
The technique developed here has been applied to the search of point-like sources in a kilometre-scale simplified detector located in the Mediterranean sea, but the results can be easily extended to other locations. We have generated 5×10^3 Monte Carlo simulated experiments, each containing the equivalent of one year of data-taking. The simulated data contains essentially a set of declination and right ascension values assumed to be the result of the final track reconstruction and selection of the detector recorded events. In order to simplify the simulation of the neutrino telescope we have assumed that the detector has a uniform response in the two local angles, that is, a flat distribution in azimuth, Φ, and in the cosine of the zenith angle, $\cos\theta$. This simplification should have a negligible impact on the comparison of the search methods, while it eases considerably the simulation of the background contribution. In the classical scheme of operation of a neutrino telescope in which up-going muon tracks are looked for, the main source of physical background comes from the atmospheric neutrinos which are isotropically distributed in the lower hemisphere. Figure 1 shows a one-year equivalent Monte Carlo data sample of atmospheric neutrinos with a Poisson mean of 2×10^4 events/year which is an estimate of the number of atmospheric neutrinos expected in a kilometre-scale detector. The upper figure shows the equatorial map of the events, while the lower figure shows the projection of the distribution in declination. Due to the Earth's rotation the distribution is uniform in right ascension. It is worth remarking that since the energy information is not used in our application, the selection of a neutrino energy spectrum is immaterial in this case.

For this work we assumed that the uncertainty in the determination of the incoming neutrino arrival direction, the so-called point spread function, does not depend on the declination. This is again an approximation which is not true for a real detector, but that should have a negligible influence on the results of our comparison. We do not consider either the dependence of the angular resolution (taken to be $\sim 0.1^\circ$) with the energy and hence no assumption about the spectral index in the signal simulation has to be made. Of course, a real data analysis must include all these effects and therefore needs a proper simulation of the neutrino detecto, buy are not critical in this work and are therefore neglected.

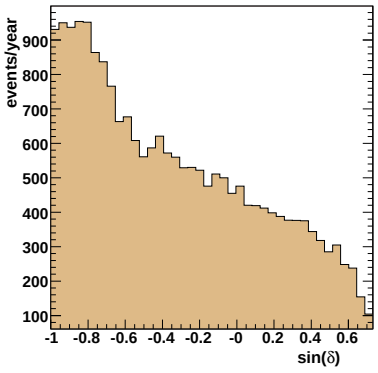
The EM algorithm in the search of neutrino point sources

The point-like source search is a clustering analysis with some special features. In our case clusters of signal events have to be spotted over a background of atmospheric neutrinos. We assume that the sources follow Gaussian distributions of the form:

$$P_S = \frac{1}{2\pi\sigma_\alpha\sigma_\delta} \exp\left(-\frac{(\alpha-\mu_\alpha)^2\cos^2\delta}{2\sigma_\alpha^2}\right) \times \times \exp\left(-\frac{(\delta-\mu_\delta)^2}{2\sigma_\delta^2}\right)\cos\delta \quad (2)$$



(a)



(b)

Figure 1: One-year Monte Carlo sample distribution of atmospheric neutrino events in (a) equatorial coordinates and (b) projected on declination.

The background distribution, on the other hand, can be easily inferred from the real data by scrambling the arrival time and azimuth angle of the events. This is a usual procedure in every searching algorithm since a knowledge of the background distribution is required to extract the significance of each cluster found by the algorithm. In our case the background distribution is estimated from the Monte Carlo simulation. The mixture model for our point-like source problem is thus:

$$p(\mathbf{x}) = \pi_{BG} P_{BG}(\delta) + \pi_S P_S(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad (3)$$

where $\mathbf{x} = (\alpha, \delta)$ is the position of the event in equatorial coordinates and $\mu_j = (\mu_\alpha^j, \mu_\delta^j)$ is the Gaussian mean that corresponds to the location of the j source. Both π_{BG} and π_S are the so-called mixing proportions of the background and source components, respectively, which satisfy the unitary condition. Note that this mixture model can be extended to the case of several sources in the sample. In this work, though, and for the sake of simplicity we will only consider the existence of one source.

The EM algorithm requires some initial values for the free parameters. The more precise these initial values, the faster the convergence will be. Hence a preclustering step, whose outcome is a set of candidate clusters, is applied. For each of these candidate clusters the EM algorithm is performed until a convergence criterion is fulfilled. As an output of the whole procedure a test statistic which reflects the goodness of the maximisation is obtained. In our case, the test used is the BIC (Bayesian Information Criterion) whose distributions, in the case of only background clusters and background plus a simulated signal, are used within the hypothesis testing theory to calculate the discovery power (or potential) of the algorithm.

Results for a 1 km^2 neutrino telescope

In this section, the results of the EM algorithm applied to the simulated data of a kilometre-scale neutrino detector are given and compared to those of a binning technique. The binning method employed is derived from the standard MACRO grid technique and is explained in detail somewhere else [4]. The discovery power is defined as the percentage of success in discovering a point-like source over the atmospheric neutrino background. Discovery is defined in turn as the finding of a cluster whose number of events is higher than that of a background fluctuation of a given number of standard deviations, usually taken to be five. Therefore, a discovery power of 50% at 5σ means that the corresponding signal will produce a fluctuation of five or more standard deviations over the background in 50% identical experiments. Figure 2 shows the power of discovery (at 5σ) for a source located at a declination of $\delta = -80°$ as a function of the mean number of observed signal events for the EM and binning algorithms. As can be seen, to reach a discovery power of 50%, the EM algorithm requires almost 20% less signal events than the bin-

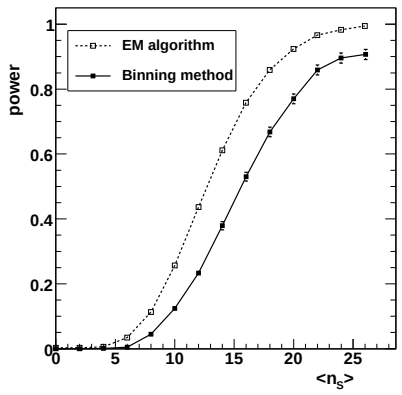


Figure 2: Discovery power as a function of the mean number of observed events from a source located at $\delta = -80^\circ$ with a confidence level of 5σ for the EM algorithm and the classical binning technique.

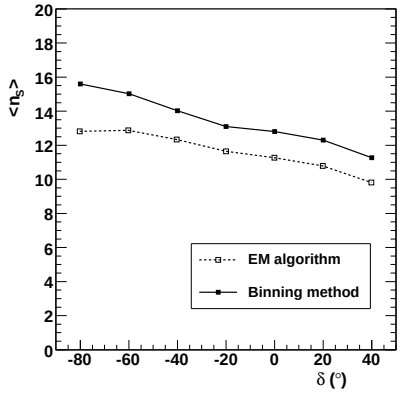


Figure 3: Mean number of selected signal events for the EM algorithm and the classical binning technique as a function of declination required to claim a discovery with a confidence level of 5σ in 50% of identical experiments.

ning technique. Although this difference changes with the discovery power, the number of events required by the EM algorithm is always smaller than that of the binning technique.

In figure 3, the mean number of signal events required by both techniques to have a discovery power of 50% (at a 5σ level) is given as a function of the declination. The EM algorithm requires between 15% to 20% less events than the binning technique to reach the same discovery power.

Summary and Outlook

In this contribution we have presented a powerful method to look for faint point-like sources based on the EM algorithm. This method has been applied to a generic neutrino telescope of 1 km^3 size located in the Mediterranean sea. This technique has been advantageously compared with the results of a classical method with binning for the same samples of simulated events. Only the muon arrival direction has been used in the algorithm, however the unbinned techniques have the advantage that other useful information such as the estimated energy of the neutrino event, can be included in the algorithm to enhance the performance of the method which cannot be done in a binning technique. Work is in progress to implement the use of the energy information in the EM algorithm.

This work is supported by Spanish MEC grants FPA2003-00531 and FPA2006-04277.

References

[1] A. P. Dempster et al. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statistical Soc. Series B*, 38:1–38, 1977.

[2] A. Achterberg et al. Five years of searches for point sources of astrophysical neutrinos with the AMANDA-II neutrino telescope. *Physical Review Letters D.*, 75:102001, 2007.

[3] *http://icecube.wisc.edu*, *http://www.km3net.org*.

[4] E. Carmona. *Study of the event reconstruction and expected performances for point-like sources of the future ANTARES neutrino telescope*. PhD thesis, Universitat de València, Spain, 2003.